**HW3: Insurance Claims Data & Analytics**

**Due: 12:00 pm, Monday Feb 25**

**Context:**

This homework lays out some analytics on the most important database in market for health, the insurance claim data. We use the publicly available All Payer Claim Data published by the State of Vermont. The databases available to us are named Vermont Uniform Hospital Discharge Data Sets. These databases consist of inpatient discharge data, outpatient procedures and services data, and emergency department data. Each data set according to the website includes:

* Case-specific diagnostic discharge data
* Some socio-demographic characteristics of the patient
* Medical reason for the admission
* Treatment and services provided to the patient
* Duration and status of the patient's stay in the hospital
* Full, undiscounted total and service-specific charges billed by the hospital

They are found in the link provided below. We want to use 2016 databases and codebooks since the 2017 upload is still too fresh and probably subject to revision.

<http://www.healthvermont.gov/health-statistics-vital-records/health-care-systems-reporting/hospital-discharge-data>

**Important reminder: Use 2016 data since the 2017 is not finalized yet.**

**Question 1: Patients’ vignette -- Files to use: 2016 Vermont Emergency Dept, Inpatient, Revenue Codes**

Write a telling story for each of the following patients using their UNIQ identifier:

**UNIQ: 507033, 40436, 859382, 1585831, 200760, 3692, 690326**

Format: A comprehensive story is one that includes narratives on patient (pt) origination (admin source), admission type, the serving hospital, pt’s demographics, insurance coverage, a nice picture of pt’s diagnoses, discharge narrative, length of stay and of course the costs, and details of the services (nature and prices) according to Revenue Codes. A good story is similar to the work of detectives: as you scan thru the rather long list of up to twenty DX codes, you want to focus on the important facts that shed some light on pt’s journey rather than just listing the DX codes one after the other with no meaningful connection between them.

Hint: The information about most of these case study pts are found in Inpatient file. However if the patient pathway begins from emergency department then you want to start there first before you move onto the IP database. As you get to the sections to talk about what is done for the patient, given his/her diagnoses, then you want to link the ED and/or inpatient files with the Revenue Code file where all the details of the services rendered to the patient are listed. In explaining major services rendered to the patient use the REVCODE variable. The REVCODE list is found in the tab named “REVCODE” of the Excel file “*HS\_2016VT\_REVCODE\_FILE\_LAYOUT\_and\_CODES*” in the same website.

Hint: Make sure you all participate in this section of the assignment. For example by assigning at least one patient to each team member. In my experience you can only learn about the complex yet exciting world of the insurance claim data and analytics, if you undertake the task yourself with minimal guidance from others. I however do not grade the team submission by individual student’s names and my suggestion is meant just as a reminder to get everybody engaged in the learning process.

**Question 2: Service and Cost Profile of Major Insurances -- Files to use: 2016 Vermont Inpatient**

Inpatient hospitalizations are identified and priced and paid by DRGs. So each hospital admission has a DRG. You can see the list of DRGs in the Excel file “*HS\_2016VT\_PUF\_FILE\_LAYOUT\_and\_CODES*” from the website. DRGs are classified in Major Diagnosis Groups or MDCs. You can find the name of the MDCs in the same Excel file. Identify the three major insurance payers in Vermont: Medicare, Medicaid and then combine all hospital admissions of the two major commercial insurances of “BLUE CROSS” + “COMMERCIAL INSURANCE” and call this combined last category as **Commercial Payers** for this question. For the 3 insurances, create a cross tabulation with the MDC categories as row heading, and the name of the 3 insurances as the column headings. The cell values of your X-tab would be the sum of the dollar value of the charges for the selected MDC for each of the 3 insurances. Turn all your dollar values to $Million and round the value to drop any decimal points so you do not overwhelm your x-tab presentation with so many multiple digits figures. Drop all non-classified, unknown, and missing value rows from your x-tab and stay focus on the known MDCs only.

After presenting your x-tab in your report draw 3 pie-charts, one per insurance, to present the graphical view of the “inpatient services portfolio” for each insurance. Each pie chart starts with the largest MDC’s share in percentages of all costs as the first slice at 12:00 o’clock.

Discuss the differences in the service portfolios across 3 pie-charts and try to relate the differences in the portfolios according to the demographics of the patients (age and sex is enough). You already know how the demographics of populations vary across the three major insurances yet you want to validate your assumptions using the real data. Your analytic writing can look like: from the pie chart of Medicaid we are observing that the top-5 MDCs are X, Y,Z,W, and Q and the reason MDC X is costing the Medicaid so much money is because Medicaid covers members with so and so demographics. However in Medicare the MDC so and so eats the largest portion of the money and that is because Medicare population is so and so .

Note: A good chunk of the grade for this question goes to your analytical story as backed by your x-tabulation figures and the pie-charts.

**Question 3: Examining the enormity of the drug use/abuse/overdose -- Files to use: 2016 Vermont ED**

Drug overdose as we learned has become a national health crisis in the US and globally. It causes many unnecessary costs and losses, financially, socially, economically, and most importantly in terms of losses of lives and/or quality of lives of the addicted individual, the family, friends, and the community. Many of us have been impacted by this crisis, one way or another, in our families, friend circles, or our communities. The governments and insurance companies are demanding the healthcare providers to pay more informational attention in submitting the drug abuse related claim data. Vermont’s state government in particular is very serious in getting as much data as possible about this health crisis from the healthcare providers and reports them to public and researchers thru data sources including the database we use here.

As we discussed in our last class, most of the drug abuse cases are brought to hospitals in a rather emergency manner so the proper database to study this question is the ED claims data. In order to study the drug abuse related cases the ICD-10 classification has assigned the entire blocks of T40 to T43 to such cases. Here are two ICD-10 code examples:

**T401X1A Poisoning by heroin, accidental**

**T424X2A Poisoning by benzodiazepines, intentional self-harm, initial encounter.**

Create a database that reports all the details form the ED file for every emergency department admission that has identified the pt with at least one drug abuse related ICD-10 code (i.e. any match with the entire code block including T40xxxx, T41xxxx, T42xxxx, and T43xxxx). Remember that these codes can appear in any location from DX1 to DX20. So do not assume that the doctor or nurse has recorded those drug use related ICD-10 codes only in the DX1 position. Also for some patients, multiple codes from the drug abuse block can be reported. This is because providers are required to report all details with regard to drug usage. However only one code from the block is enough for you to identify the patient as a drug use/abuse ED visit. If your data retrieval codes (SQL or any other search code) are designed and implemented correctly, you will see slightly over 2,000 cases of ED admissions for drug use/abuse. If you are seeing fewer or more please debug your code and run again.

Once you have created the drug use/abuse analytical database answer these questions:

* How many ED visits exactly have been diagnosed as drug user/abuser?
* There is a myth that the drug use/abuse has been a male problem and that women have much better protection measures to stay away from drug use/abuse let alone overdoses severe enough that lead to an ED admission. Can you check if your data supports this gender bias myth?
* Tens of millions of dollars reportedly were spent on drug use related cases that year alone. Can you find the exact dollar amount for your identified patients in this question? Of the three insurances in Question 2, what was share of each of the total payments?
* Recent breakthroughs in the dark side of chemistry of drug development have done nothing but damage to humanity. The use of synthetic narcotics is rising alarmingly in part due to the marketing campaign for such meds. On the other front the public is ill informed of the danger of those new generation of lab created drugs that are supposedly improving brain’s performance ([read more here](https://supportsystemshomes.com/silicon-valley-high-functioning-addict-treatment/) or [here](https://www.wbur.org/hereandnow/2018/08/24/drug-use-mainstream-silicon-valley)). Use the ICD-10 codes of T404xxx and T4362xx to identify only a small sample of such patients. How many of patients have been brought to ED for diagnosis related to synthetic narcotics or amphetamines?
* Name the 3 zip code regions with the highest numbers of drug use/abuse cases.
* What are the 10 most common diagnoses of drug use/abuse?

**Hint:** Remember that for this last part of this question you can have multiple drug use/abuse ICD-10 codes for some patients. See the patient with UNIQ = 19314 for example. He’s been diagnosed with two drug use/abuse codes of T40605A and T4275XA, among a long list of other diagnoses. So if you pool all the reported codes you must get the number of all reported codes for all patients higher than the number of ED visits. Again, you need at least one code to find the patient to be a user/abuser, but then you need to pool all codes for all patients to answer this section of the question.